Transportation Research Part C 141 (2022) 103758

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

A Spatio-Temporal autocorrelation model for designing a carshare system using historical heterogeneous Data: Policy suggestion

Zesheng Cheng [a,b], Taha Hossein Rashidi [a,*], Sisi Jian [c], Mojtaba Maghrebi [a,d], Steven Travis Waller [a], Vinayak Dixit [a]

[a] Research Centre for Integrated Transport Innovation (rCITI), School of Civil and Environmental Engineering, UNSW, Sydney, Australia
[b] College of Computer Science and Technology, Qingdao University, Qingdao, PR China
[c] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong Special Administrative Region
[d] Department of Civil Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

ARTICLE INFO

Keywords:
Car-sharing
Spatial-temporal Auto-correaltion
Machine Learning
Social Media Data
Policy Analysis

ABSTRACT

As an emerging urban mobility service, carsharing has become increasingly popular worldwide. To understand customers' needs and optimize the design of service networks, the usage of car-sharing vehicles whose trips are recorded by the operators has been applied in research estimating carsharing demand. However, as a form of spatio-temporal correlated data, the underlying spatio-temporal information included in such carsharing records has not been investigated in existing models of carsharing demand. Meanwhile, due to the supply limitation of carsharing stations, some demand cannot be fulfilled and thus remains unrecorded in the operational data. Unrealized demand may lead to underestimation of carsharing demand and therefore an incorrect vehicle deployment strategy by the service providers. In view of these issues, this paper develops an innovative approach to estimating the actual demand at a carsharing station with operational data from *GoGet*, the largest carsharing company in Australia. The accuracy of the estimation is improved by adding spatio-temporal correlated variables as well as variables from emerging data sources such as social media. To explore the latent space-and-time correlated information, spatio-temporal autoregressive and moving-average models have been applied. Based on the results of the analysis, the paper also provides recommendations related to the operation policies of the service providers.

1. Introduction

With the advent of technologies such as geographic information systems (GIS), remote sensing (RS), and global positioning systems (GPS), a large amount of spatio-temporal data is being utilized in the field of transport engineering for purposes of predicting travel demand, estimating traffic status, and exploring underlying traffic dynamics. These spatio-temporal data are mostly extracted and stored discretely in terms of time and space, and always present complex space-and-time correlations (Yao, 2003). Previously, due to the lack of effective modelling tools, it has been challenging to analyze spatio-temporal data quantitatively and explore the latent

* Corresponding author.
E-mail addresses: czs_110@hotmail.com (Z. Cheng), rashidi@unsw.edu.au (T.H. Rashidi), cesjian@ust.hk (S. Jian), maghrebi@unsw.edu.au (M. Maghrebi), s.travis.waller@gmail.com (S.T. Waller), v.dixit@unsw.edu.au (V. Dixit).

https://doi.org/10.1016/j.trc.2022.103758
Received 29 June 2020; Received in revised form 18 May 2022; Accepted 8 June 2022
Available online 18 June 2022

information contained within it. Grappling with these issue, Curry was the first to present a preliminary definition of spatio-temporal analysis (Curry, 1970). He suggested that spatio-temporal dynamic processes can be regarded as the spatial expansion of time series. Subsequent work on spatio-temporal analysis built on Curry's approach. Thus, Cliff and Ord (Cliff and Ord, 1975) demonstrated the conception of and an effective modelling structure for spatio-temporal autocorrelation. This framework referred to spatial analysis methods in geography and combined them with analyzes based on time series. Using this framework, Martin and Oeppen (Martin and Oeppen, 1975) proposed a universal approach to modelling spatio-temporal autocorrelation. They proposed a space–time autocorrelation function and a partial autocorrelation function to quantify the spatio-temporal autocorrelation. By means of these two functions, researchers could preliminary determine the structure of a spatio-temporal autocorrelation model. This general approach has then been applied to a variety of fields, including economics (Lv et al., 2021; Pfeifer and Deutsch, 1980), criminology (Pfeifer and Deutrch, 1980), and hydrology (Cressie and Majure, 1997, Deutsch and Ramos, 1986).

In this study, we focus on applying a spatio-temporal data analysis approach to carsharing demand modelling. As an emerging urban mobility service, carsharing has become popular worldwide. One of the common carsharing mode is "go and get". Under this mode, "stations" are generated around the serviced regions by the TNC and users could rent and return provided vehicles in different stations. Therefore, optimizing the vehicles distribution at different stations are worth to be discussed. To understand customers' needs and optimize the design of service networks and the allocation of resources, carsharing operators collect abundant spatio-temporal operational data concerning the usage of carsharing vehicles and the travel patterns of carsharing customers. The aim of the study is to discuss the application and explore the latent information from the historical records.

2. Related work

Spatio-temporal autocorrelation analysis approaches have been applied to various transport-related problems. For example, Kamatianakis and Prastacos (Kamarianakis and Prastacos, 2005) used a space–time autoregressive integrated moving average (*STARIMA*) model to forecast short-term traffic flow. This study used correlated spatial and temporal variables to describe the condition of the links in the network as well as the relationship between them. By applying these variables, instead of modelling different links separately, traffic-flow estimation of the complex network can be simplified to a single model. The *STARIMA* model thus served as a potential bridge between traffic-flow equilibrium theories and real-world conjectures. Dong and Cirillo suggested that commuting departure times can be selected dynamically according to spatial and temporal information. Simulation results in their paper showed that by applying correlated spatial and temporal variables in schedule planning, the commuting services provided were 15% closer to the real demand. Combining the space–time model with a data-linkage approach, the study concluded that the planned schedule satisfied the requirements of both passengers and the service providers. Yang et al. (2019) presented a deep-learning approach to estimate real-time parking occupancy with multiple spatio-temporal data sources. They captured the latent spatio-temporal correlation between parking, traffic, and weather, and provided an accurate prediction of parking occupancy in target regions. When used to analyze a case study, compared with baseline models without correlated spatial and temporal data, the model described in the paper was 42% more accurate. These studies have demonstrated the fruitfulness of incorporating spatio-temporal autocorrelation analysis into research on travel demand and traffic dynamics modelling.

Researchers have utilized these spatio-temporal data and developed discrete-choice models and statistical models to explore the behaviour of carsharing customers (e.g., (Dong et al., 2018, Celsor and Millard-Ball, 2007, Cepolina and Farina, 2012, Cervero et al., 2007, Ciari et al., 2013, Costain et al., 2012, De Lorimier and El-Geneidy, 2013, Fagnant and Kockelman, 2014, Firnkorn and Müller, 2011, Habib et al., 2012, Morency et al., 2012, Morency et al., 2011, Schaefers, 2013, Schmöller et al., 2015, Jian et al., 2016, Jian et al., 2017)). Detailed literature reviews on carsharing behavioural modelling can be found in (Jorge and Correia, 2013, Jian et al., 2016). As concluded by (De Lorimier and El-Geneidy, 2013, Jian et al., 2016, Jian et al., 2017), customers' demand for carsharing is significantly affected by the location of their origin, the location of carsharing vehicle, and when they need to travel. These results highlight the impacts of space and time on carsharing demand; however, the underlying spatio-temporal autocorrelation has not yet been investigated in existing models of carsharing demand. In fact, the usage of a carsharing vehicle might be influenced by the usage of past time-slots and other vehicles near to the target vehicle. Hence space–time information plays an important role in developing a viable operational strategy for the service providers.

Furthermore, another issue requiring additional research is the underestimation of demand when the observed demand is equal to supply. Since the travel demand observed from historical data can never exceed the supply, the maximum number of booked trips that can be observed from operational data cannot exceed the number of available carsharing vehicles. However, the actual demand can be higher than the supply. But due to the supply limitation, some demand cannot be fulfilled. This unrealized demand cannot be directly estimated using trip records. The underestimation of carsharing demand can lead to incorrect vehicle deployment strategies, which will further impair the profitability and level-of-service of carsharing systems.

Therefore, exploring latent demand can help TNC and researchers to gain a more practical understanding towards the relationship between supply and demand of the services. Relevant problems have been discussed in various transport fields, such as public transport demand estimation (Lin, 2017, Codecá et al., 2017), bike-sharing (El-Assi et al., 2017) and automats vehicle requirement evaluation (Boesch et al., 2016; Zhang et al., 2019). Meanwhile, different approaches have been applied to solve those problems including comprehensive spatial analysis (El-Assi et al., 2017; Lv et al., 2021), information hypothesis (Lin, 2017) and network simulation (Codecá et al., 2017, Boesch et al., 2016). Yet for carsharing, appropriate approach to explore the latent demand has not been adequately addressed in the literature. To this end, this study is motivated by the following research questions:

- What is the underlying spatio-temporal correlation of carsharing customer demand?

- How can we exploit latent carsharing demand using observed trip-booking records?

With the global emergence of shared mobility options, the classical mobility types of private car ownership, public transportation and non-motorized options are being replaced by multimodal mobility-as-a-service and shared mobility options. In societies with numerous shared mobility options, private mobility provers and government agencies must take advantage of the wide range of emerging data sources and appropriate modelling approaches to analyze various sources' massive amounts of information. The main contribution of the paper is summarized as:

- The paper presents a large set of seemingly unrelated data sources (including emerging types and classical types). It provides an overview of technical methods to analyze such a massive amount of data of different types.
- The paper provides an extremely timely approach for shared mobility providers and planners to develop understating about information resources and technical methods on how they can picture demand for their service. Demand prediction in this context might be done daily rather than long-term planning as travel patterns shift day-today.
- With extensive social data (including social-demographic and social media data), the paper demonstrates how classical data sources can be complemented to provide a holistic view of the demand for mobility services using advanced modelling methods.
- By estimating the realistic demand for carsharing stations, the paper attempts to point out and conclude the factors that may influence the operation situation and user selection of carsharing service located at a different place according to heterogeneous historical data.

Following sections of the paper is organized as follows: The variables used in the study as well as a data exploration report is presented in Section 3. Detailed modelling framework will be discussed in Section 4. Section 5 is the process of coefficient calculation then demonstrate the estimation results. Section 6 analyzes the estimation result and provides operation policy to the service provider according to the analysis. Finally, Section 7 concludes.

3. Data Description

This section describes the data and variables used in the paper then explore and report the latent relationship between independent variables.

3.1. Raw data and Pre-processing

Raw data in this study derive from the following three main data sources:

1) **GoGet Car Renting Record**

The *GoGet* CarShare company operates Australia's first and largest carsharing network (GoGet, 2020). To facilitate our research, the company provided car-renting records from its carsharing network for the period from 1st January 2017 to 30th April 2017, a total of 120 days, and their corresponding vehicle and station information around the world. In the model developed in this study, the dataset is split to first 90 days and last 30 days and applied as training set and testing set respectively. The raw data encompassed 1,714 stations and more than 4,000 vehicles. Considering the convenience of collecting land-use and socio-demographic data, only stations located in Sydney, NSW. Meanwhile, stations located in a range of 250 m will be regarded as one station. After data cleaning, there is a dataset containing information about a total of 945 stations and 2454 vehicles.

2) **Land-use and Socio-demographic Data**

Land-use and socio-demographic data are collected based on a mesh block system developed by the Australian Bureau of Statistics (ABS). The system divides Australia into more than 35,000 mesh blocks. In metropolitan areas with higher population density, the average area of mesh blocks is around 1 km^2. The ABS also provides information about each mesh block, including population, area, and land use. We discuss this information in more detail below (Statistics, 2016a, Statistics, 2016b).

3) **Social Media Data from Foursquare**

We also used an application programming interface (API) to collect public information from Foursquare. Because landmarks registered on Foursquare have attached geo-locations, the API can be used to identify the top 50 points—that is, the 50 places where the most users check in—as well as the total number of check-ins at each of those points. The sum of these check-ins represents, to some extent, the area's ability to attract people. Therefore, we use both the number of users and the number of check-ins as independent variables in our model.

3.2. Dependent variables

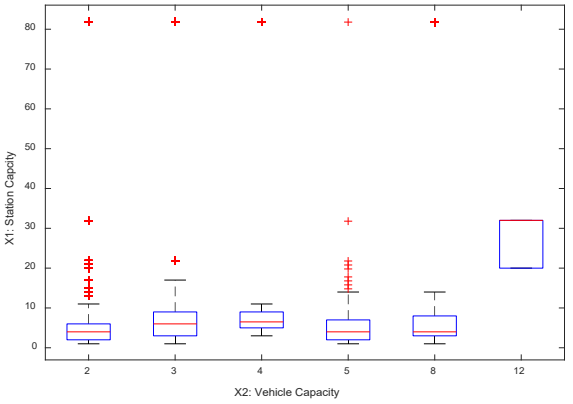
Providing many under-utilized vehicles above the realized demand will reduce the usage rate of vehicles. For example, if there are

three vehicles, A, B and C, serving at a target station Z, where only two vehicles are used every day during the estimation period (A and B are used on day 1, B and C are used on day 2, and A and C are used on day 3), then the realised demand of the station can be satisfied by providing only two vehicles to station Z. It means that station Z provides one under-utilized vehicle above the realised demand. However, the real-world situations might be rather different. In some timeslots, all 3 vehicles may be used, while in other timeslots, none of them is used. Therefore, to define whether under-utilized vehicles have been provided at a station, an ad-hoc yardstick (95% for example) has been utilized in this paper. If at least one vehicle has not been used in more than 95% of its available timeslots according to the historical record, the station seems to provide under-utilized vehicles. The defined yardstick can be set flexibly according to the operational strategy. It should be approximately equal to the idle-time rate of a vehicle which might be acceptable to TNC providers.

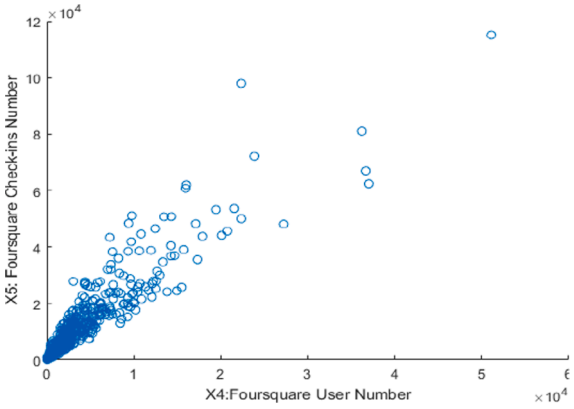
According to this, whether a target vehicle has been used at least once per day is defined as the dependent variable in this paper. Once the usage condition of the provided vehicles is known, the realistic demand can be estimated by the process demonstrated in Fig. 1 below. Setting this dependent variable, as noted, makes the model a binary classification model, in which "1″ stands for the vehicle being used and "0" otherwise. Therefore, the dependent variable is a vector with 2454 (the total number of vehicles available), * 90 (days) binary terms in the training dataset. While the dependent variable is a vector with 2454 (the total number of vehicles available) * 30 (days) binary terms in the testing dataset. For example, if a target vehicle has been booked once on day 1, twice on day four and leisure on day 2, 3 and 5. Then the input sequence would be [1(used once), 0, 0, 1(used at least once), 0].

3.3. Description of independent variables

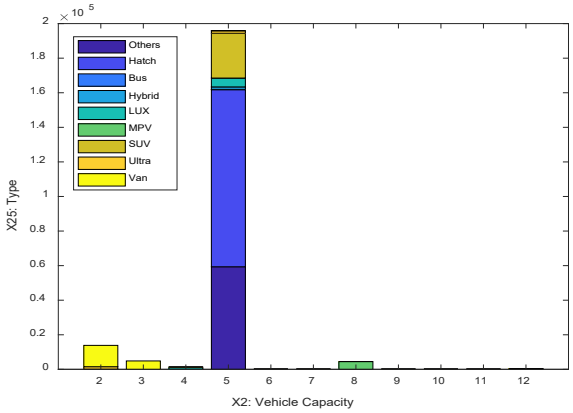
In addition to the spatial and temporal variable mentioned previously, after data pre-processing, we used one dependent variable and 23 independent variables in our analysis. Table 1 divides these variables into three types and presents a detailed description for each of them.



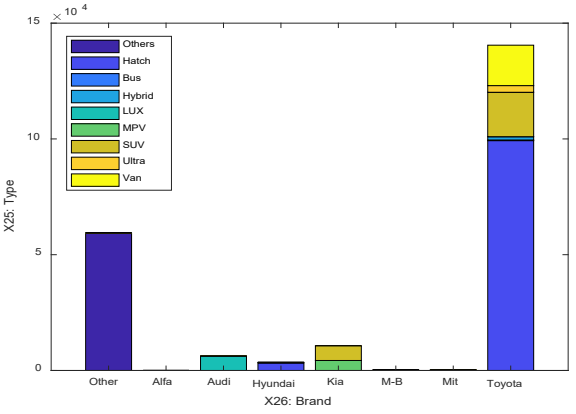
a): Vehicle Capacity vs. Station Capacity



b): User Number vs. Check-ins Number



c): Vehicle Number vs. Type



d): Brand vs. Type

Fig. 1. Notable correlation between variables.

3.4. Data exploration

Data exploration is commonly used both to test the quantity of data and to improve its quality by 'cleaning it up'. In this paper, bivariate analysis is used to test the relationship between every pair of independent variables. Theoretically, if there is a strong correlation between two independent variables, one of them can be excluded from the model to increase its efficiency.

The results of data exploration for the 27 variables listed in Table 1 show a slight level of correlation between variable pairs, which is mostly smaller than 0.1. However, the correlations between some pairs of variables are considerable. Four pairs from 729 (27*27) have been selected and their correlations are shown in Fig. 1(a)-Fig. 3(d). Amount them, Fig. 1(a) displays the outliers in the dataset. Fig. 1(b)-Fig. 3(d) demonstrate different noteworthy correlation between variables including linear and one-to-one relationship. They are also the largest and only three pairs of variables which have correlation larger than 0.3. Paragraphs below discuss the methods to apply those variables with considerable correlations.

In Fig. 1(a), it is worth noting that there are two outliers with respect to station capacity (variable 2). These two outliers represent stations providing 82 and 34 vehicles which are the largest two among all stations. Meanwhile, these two stations are also the busiest stations according to the historical record. Although outliers are usually removed to ensure the performance of the model, large amount of information will loss in our case if excluding these two stations from the dataset. Therefore, the two outliers are kept in the dataset.

In Fig. 1(b), a linear relationship between Foursquare user numbers and the total number of check-ins is identified. The correlation between these two variables can be numerically measured by using the following Pearson correlation formula:

$$correlation(X,Y) = \frac{covariance(X,Y)}{Var(X)*Var(Y)} \tag{1}$$

The calculated correlation value is 0.945, which shows a strong positive linear correlation. This result suggests that one of the

Table 1
Variable Description.

Index	Name	Description	Range	Source
Dependent Variable				
y	Usage	Whether the target vehicle has been used during a given time period	[1,0]	*GoGet*
Independent Variables – Target mesh block, station, vehicle				
$\mathbf{x^1}$	Station Capacity	How many vehicles are available at the station to which the target vehicle belongs	1–82	*GoGet*
$\mathbf{x^2}$	Vehicle Capacity	How many seats in the target vehicle	[2,3,4,5,8,12]	*GoGet*
$\mathbf{x^3}$	Commercial Area Ratio	The proportion of different types of land use in total area of the target mesh block	0–1	ABS
$\mathbf{x^4}$	Educational Area Ratio			
$\mathbf{x^5}$	Industrial Area Ratio			
$\mathbf{x^6}$	Parking Land Area Ratio			
$\mathbf{x^7}$	Rural Area Ratio			
$\mathbf{x^8}$	Transport Area Ratio			
$\mathbf{x^9}$	Population	Population of target mesh block	0–5524	ABS
$\mathbf{x^{10}}$	Area	Area of target mesh block	0.09–906 km^2	ABS
$\mathbf{x^{11}}$	Brand	Brand of target vehicle	[Alfa, Audi, Hyundai, Kia, M−B, Mit, Toyota, Others]	*GoGet*
$\mathbf{x^{12}}$	Type	Type of target vehicle	[Hatch, Bus, Hybrid, LUX, MPV, SUV, Ultra, Van, Others]	*GoGet*
Independent Variables – Information about Adjacent Spaces and Times				
$\mathbf{x^{13}}$	Nearest Distance	The distance between nearest vehicle of same brand and type as target vehicle	0–20 km*	*GoGet*
$\mathbf{x^{14}}$	Time not been used	How long has the target vehicle not been used before this time period	0–88 day	*GoGet*
$\mathbf{x^{15}}$	Use Rate / Month	The usage rate of the target vehicle in the past 30 days	0–1	*GoGet*
$\mathbf{x^{16}}$	Surrounding Commercial Area	The proportion of different types of land use in total are of mesh blocks around	0–1	ABS
$\mathbf{x^{17}}$	Surrounding Educational Area			
$\mathbf{x^{18}}$	Surrounding Industrial Area			
$\mathbf{x^{19}}$	Surrounding Parking Land Area			
$\mathbf{x^{20}}$	Surrounding Rural Area			
$\mathbf{x^{21}}$	Surrounding Transport Area			
Independent Variables – Social Media Data from Foursquare				
$\mathbf{x^{22}}$	Foursquare User Number	How many users check in in the target mesh block	0–51097	Foursquare
$\mathbf{x^{23}}$	Number of Check-ins	Total number of check-ins in the target mesh block	0–115092	Foursquare

(*If a vehicle of the same type and brand cannot be found within 20 km, then the value of the variable will be written as 100 km).

variables should be excluded from the model. Therefore, user number is selected to be applied in the model alone because it has smaller order of magnitude and variance.

Fig. 1(c) demonstrates the relationship between two categorical variables—namely, vehicle capacity and car type. Spearman's rank correlation coefficient (Sedgwick, 2014) has a similar theoretical basis with commonly used Pearson correlation. Further, it is less sensitive to outliers and can be used to measure the correlation between two categorized variables. The Spearman's rank correlation coefficient of vehicle capacity and car type is calculated as 0.3163. Since this Spearman's rank correlation coefficient suggests that a weak correlation between these variables, both can be included in the model.

Fig. 1(d) shows the relationship between vehicle types and vehicle brands in the context of *GoGet's* carsharing services. The correlation between them are 0.680 and there are many one-to-one relationships, meaning that for some brands there are only vehicles of one type. For other brands, there are only a few types of vehicles. Therefore, the data for vehicle types and brands are aggregated as shown in Table 2.

4. Methodology

Fig. 2 elaborates the developed process of the model by this paper and demonstrates the relationship between the model's different components, including the observed demand from historical data and the unrealized demand, which is the missing part of the historical record under different scenarios. To estimate the realistic demand of the carsharing service, if there are one or more under-utilized vehicles at one station, the number of vehicles required to satisfy the realistic demand can be estimated by excluding under-utilized vehicles. If there are no under-utilized vehicles, vehicles can be added one by one to the station of interest to estimate the realistic demand. The developed model can determine the service condition and whether it is over-utilized. Then, the realistic demand of the station is satisfied by the highest number of vehicles before over-utilizing occurs.

The process of creating a *STARMA* model includes structure identification, coefficient calculation, and model validation. The rest of Section 4 explains each step in turn following a brief introduction on spatial and temporal variables.

4.1. Related spatial and temporal variables

In a *STARMA* model, the observed (dependent) variable *Yi(t)* collected at zone (a small region of interest) i and timeslot t might be correlated with 3 different types of spatio-temporal variables, which are temporal lag variables, temporal lag random disturbance terms and spatial autocorrelation variables:

1) **Temporal lag variables: $Y_i(t-1)$, $Y_i(t-2)$, …, $Y_i(t-k)$**

Variables of this type refer to the observed values collected at zone *i* and timeslot *(t-1)* to timeslot *(t-n)*. $\boldsymbol{Y_i(t-k)}$ in equation (1) below named as n^{th}-order spatial lag of $\boldsymbol{Y_i(t)}$.

2) **Temporal lag random disturbance terms: $\varepsilon_i(t-1)$, $\varepsilon_i(t-2)$, …, $\varepsilon_i(t-l)$**

$\boldsymbol{\varepsilon_i(t)}$ is a normally distributed random disturbance term in the regression process of $Y_i(t)$. Temporal lag random disturbance terms, which is presented by $\boldsymbol{\varepsilon_i(t-l)}$ in equation (1) below, are used to eliminate the influence of periodic or large fluctuations in the variables.

3) **Spatial autocorrelation variables: $W^1Y_i(t)$, $W^2Y_i(t)$, …, $W^nY_i(t)$**

$\boldsymbol{W}$ is the spatial weight matrix which reflects the 'adjacent' relationship between two zones. For example, if zone *i* is adjacent to zone *j*, then $W_{ij} = 1$, and 0 otherwise. Then the matrices are normalized to substitute into operations. After normalization, if zone *i* is adjacent to three other zones, the value of all the related positions in the matrix is 1/3. In equation (2) below, $\boldsymbol{W^n}$ stands for n-order adjacent matrix.

Basically, there are two ways to geographically define adjacency. The first way is by determining whether there are common edges or corners between the two zones. The second way is by determining whether the distance between the centroids of two areas is smaller than a pre-set value. According to Tobler's first law of geography (Tobler, 1970), 'everything is related to everything else, but near things are more related than distant things'. In accordance with this intuition, the spatial weight matrix reflects, to some extent, the interaction of observed values between areas. In this paper, if vehicles with same brand and type are located in a range of 0.5 km, they will be defined as 1st-order adjacent. While those located in a range of 0.5–5 km are defined as 2nd-order adjacent. Fig. 3 shows two examples of target stations (station 1 and 2) as well as their 1st-order spatial adjacent station (marked by yellow) and 2nd-order spatial

Table 2
Aggregated Vehicle Brand and Type Variables.

Index	Name	Description	Range	Source
Independent Variables – Information about Adjacent Spaces and Times				
X^{11}	Brand	Brand of target vehicle	[Audi, Hyundai, Kia, Toyota, Others]	*GoGet*
X^{12}	Type	Type of target vehicle	[Hatch, Hybrid, MPV, SUV, Ultra, Van, Others]	*GoGet*

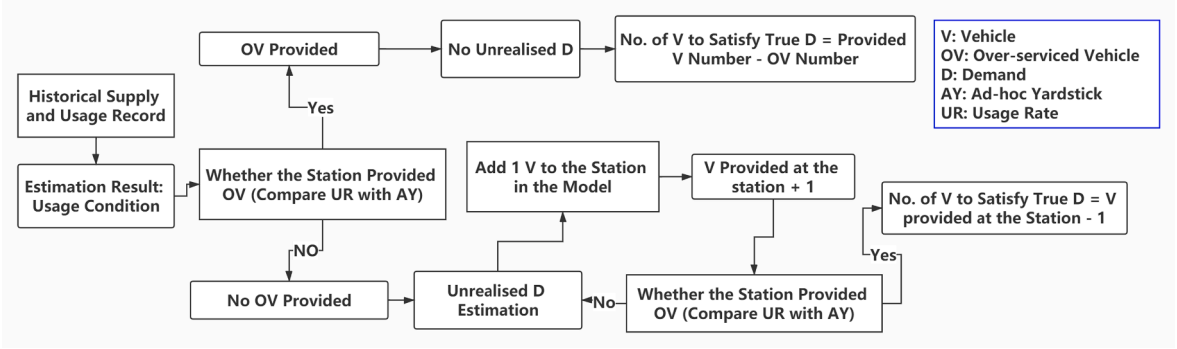


Fig. 2. Flow chart of the developed approach.

adjacent station (marked by green). The radius of the marks stands for the value of normalized spatial weighted of them respectively. There are other methods to defined adjacency, such as functionality similarity matrix and mobility pattern similarity matrix as well. Detailed discussions of them will be shown in Appendix A.

The structure identification process introduced in section 4.2 below presents a method for selecting appropriate variables. If the model contains a combination of all three of the variable types just described, then it becomes a *STARMA* model. The general form of a *STARMA* model can be represented as followed (Pfeifer and Deutrch, 1980):

$$Y_i(t) = \sum_{k=1}^{p} \sum_{h=0}^{m_k} \varphi_{kh} L^{(h)} Y_i(t-k) - \sum_{l=1}^{q} \sum_{h=0}^{m_l} \theta_{lh} L^{(h)} \varepsilon_i(t-l) - \varepsilon_i(t) \tag{2}$$

Where:

$Y_i(t)$ Observation value at zone i, time t.

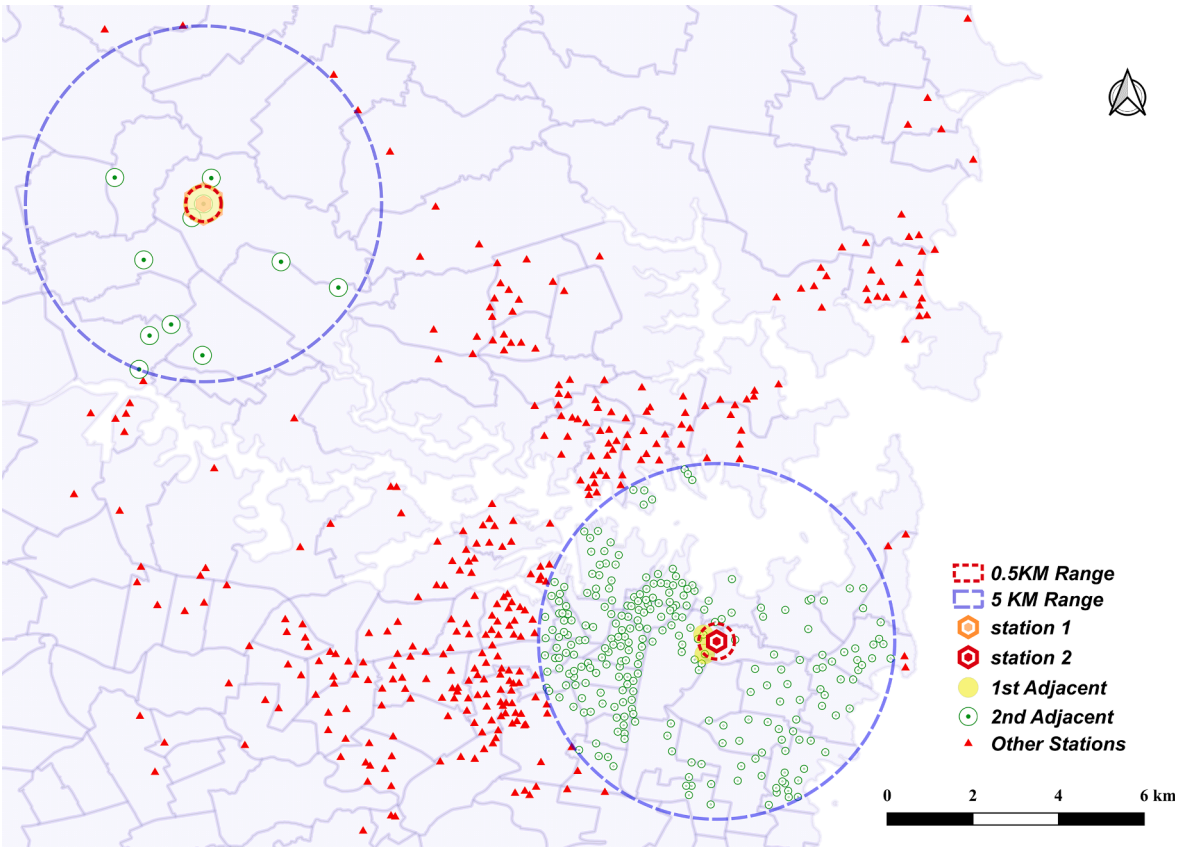


Fig. 3. Example of spatial correlation.

$Y_i(t\text{-}k)$ k^{th}-order temporal lag variable at zone i.
$\varepsilon_i(t\text{-}l)$ l^{th}-order random disturbance term at zone i.
p Largest autoregressive order.
q Largest moving average order.
m_k Largest spatial autocorrelation order of k^{th}-order temporal autoregressive term.
m_l Largest spatial autocorrelation order of l^{th}-order temporal moving average term.
φ_{kh} Coefficient of k^{th}-order temporal autoregressive, h^{th}-order spatial correlated term.
θ_{lh} Coefficient of l^{th}-order temporal moving average term, h^{th}-order spatial correlated term.
$\varepsilon_i(t)$ Normal distributed random error term at zone i time t.
In the function, $L^{(h)}$ is an h^{th}-order spatial correlated operator:

$$L^{(h)}z_i(t) = \sum_{j=1}^{N} w_{ij}^{(h)} z_j(t) \tag{3}$$

Where:
$w_{ij}^{(h)}$ Term representing whether a zone is h^{th}-order adjacent to zone j in h^{th}-order spatial weight matrix $\boldsymbol{W}^{(h)}$.

4.2. Structure of Spatio-temporal autocorrelation

As mentioned previously, the specific structure of the model can be determined by the space–time autocorrelation function and space–time partial autocorrelation function.

The autocorrelation coefficient of the h^{th}-order spatial correlation and k^{th}-order temporal lag can be calculated as followed (Martin and Oeppen, 1975):

$$\rho_h(k) = \frac{Covariance(\left[\boldsymbol{W}^{\boldsymbol{h}}Y(t)\right], \left[\boldsymbol{W}^{0}Y(t-k)\right])}{\sqrt{Variance(\boldsymbol{W}^{\boldsymbol{h}}Y(t)) \bullet Variance(\boldsymbol{W}^{0}Y(t))}} \tag{4}$$

Where:
$\rho_h(k)$ Space-time autocorrelation coefficient at h^{th}-order spatial correlation and k^{th}-order temporal lag.
$Y(t)$ Value of dependent variable Y at time t for all zones.
$Y(t\text{-}k)$ Value of dependent variable Y at time t-k for all zones.
$\boldsymbol{W}^{\boldsymbol{h}}$ h^{th}-order spatial weight matrix.
$\boldsymbol{W}^{\boldsymbol{0}}$ Zero-order spatial weight matrix, a square unit matrix reflecting that all zones are zero-order adjacent with themselves.

The space–time partial autocorrelation coefficient can be calculated by the Yule-Walker equations system as related to the autocorrelation coefficient; the procedure is shown in Eq. (4) (Martin and Oeppen, 1975):

$$\rho_h(k) = \sum_{t=1}^{M_T} \sum_{l=1}^{L_k} \varphi_{kh}\rho_{l-h}(t-k) \tag{5}$$

Where:
$\rho_h(k)$ Space-time autocorrelation coefficient at h^{th}-order spatial correlation and k^{th}-order temporal lag.
φ_{kh} Space-time partial autocorrelation coefficient at h^{th}-order spatial correlation and k^{th}-order temporal lag.
M_T The max order index of temporal lag.
L_k The max order index of spatial correlation at k^{th}-order temporal lag.

Both autocorrelation coefficients and partial coefficients may present two kinds of trends in the dimensions of both space and time. These two trends are:

1) truncation: the absolute value of the coefficient after p^{th}-order suddenly decays to approximate 0 and remains 0 after p;
2) trailing: the absolute value of the coefficient keeps decreasing or fluctuating but always remains larger than 0.

By considering the combination of these two trends, the structure of the model can be preliminarily determined (Martin and Oeppen, 1975). If both coefficients are trailing, then the model should be a *STARMA* model.

4.3. Coefficient calculation

After knowing the relevant spatio-temporal variables and the structure of the model, the next step is to evaluate the coefficient of each variable in the model. Seven machine learning methods including *Linear Method with Maximum Likelihood Estimation (L-MLE), Binary Logistics Model (BL), Classification and Regression Tree (CART), Support Vector Machine (SVM), Naive Bayesian Classification (NBC), Random Forest (RF), Extreme Gradient Boosting (XgBoost)*, have been applied in this paper. In the study, those methods are mostly trained and tested by MATLAB. The rest of the subsection provides a brief introduction and the details about tuning the parameters in MATLAB for each method.

4.4. Coefficient calculation

After knowing the relevant Spatio-temporal variables and the structure of the model, the next step is to evaluate the coefficient of each variable in the model. Seven machine learning methods, including Linear Method with Maximum Likelihood Estimation (L-MLE), Binary Logistics Model (BL), Classification and Regression Tree (CART), Support Vector Machine (SVM), Naive Bayesian Classification (NBC), Random Forest (RF), Extreme Gradient Boosting (XgBoost), have been applied in this paper. In the study, those methods are primarily trained and tested by MATLAB. The rest of the subsection provides a brief introduction and details about tuning the parameters in MATLAB for each technique.

1) **Linear Method with Maximum Likelihood Estimation**

L-MLE is a commonly used linear algorithm applied to calculate the coefficient of STARMA models (Cliff and Ord, 1981; Pfeifer and Deutrch, 1980) in geography and other fields. The core approach of the L-MLE algorithm is to calculate the most likely value of coefficients based on the standard joint normal distribution of the error term εi(t). In MATLAB, the L-MLE algorithm can be applied using the build-in function [fitglm]. The split value, which is the unique tunable parameter, is set as the default value of 0.5. Only the predicted value higher than 0.5 will be regarded as 1 (used at least once during the target time).

2) **Binary Logistic Model**

Multinomial Logistic Model (MNL) is a commonly used classification method based on a logistic regression transform from linear regression (Engel, 1988). BL model is a particular form of the MNL. The coefficients of this type of model can be obtained by finding a set (β0, β) that maximizes:

$$L(\beta_0, \boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{e^{O_i(\beta_0 + \boldsymbol{\beta x_i})}}{1 + e^{(\beta_0 + \boldsymbol{\beta x_i})}} \tag{6}$$

Where:
O_i i-th binary observation value in the dependent variable.
$\boldsymbol{x_i}$ A vector of explanatory (independent) variables of i-th independent variable.
β_0 The constant term in the model.
$\boldsymbol{\beta}$ A vector of coefficients, in one-to-one correspondence with the explanatory variable.
When applied in MATLAB, no specific parameter tuning is required for the BL model.

3) **Classification and Regression Tree**

The basic idea of the CART algorithm is to divide a given space into a set of rectangular areas and then fit the points in each area to a constant or to a simpler model (Breiman and Ihaka, 1984). In MATLAB, the CART algorithm can be applied using the build-in function [fitctree]. By setting the parameter "OptimizeHyperparameters" of that function as "auto", the process will automatically optimise the classification tree by pruning and tree depth controlling based on cross-validation loss.

4) **Support Vector Machine**

SVM applies hinge loss to calculate the empirical risk. The decision boundary of an SVM model is the maximum-margin hyperplane of training examples. With the kernel method, SVM is identified as a stable sparse classifier that can be applied to solve linear and non-linear classification problems (Noble, 2006).

In MATLAB, SVM can be applied by using the in-build function [fitcsvm]. A polynomial kernel has been used as the kernel function of the SVM model in the study. Meanwhile, the hyperparameters are automatically optimised by the in-build function, according to a 5-fold cross-validation loss.

5) **Naive Bayesian Classification**

NBC is a simple probability classifier that applies Bayes' Theorem to train the model. It assumes that all the dataset features are strongly independent of each other (Leung, 2007). In MATLAB, NBC can be applied by using the in-build function [fitcnb]. When applied in MATLAB, no specific parameter tuning is required for the NBC model.

6) **Random Forest**

RF is an ensemble process of CART. It applies an ensemble learning technique that uses a bagging algorithm to integrate several CARTs (Hastie et al., 2009). These CARTs are independent of each other, and the estimates of the forest are determined by their voting and mode. In MATLAB, the RF algorithm can be applied using the build-in function [TreeBagger]. After several tests, considering both accuracy and computing cost, the number of trees in the model is 500. In contrast, each tree is trained by a maximum of six randomly

selected variables in the study.

7) **Extreme Gradient Boosting**

XgBoost is another widely used ensemble learning method based on CART. Essentially, the model creates new trees continuously to fit the residuals of the previous steps method and then sums the result of each tree up as the final output (Chen et al., 2015). In the study, XgBoost was applied in Python manually. Relevant parameters such as learning rate, max depth of trees and minimum child wright have been tuned by Grid Search method using library [GridSearchCV].

4.5. Model validation

L-MLE and *BL* can be validated through a hypothesis test, which is commonly used in parametric modelling. By examining the *p-value* of the coefficient of each independent variable, whether the variable is significant or not in the model can be determined.

The importance of the variables in other five models can be tested following the steps below. This method is called out-of-bag estimation of feature importance, or the *OOB-factor* (Hastie et al., 2009).

1) Compute the root mean squared error (RMSE) for the estimation result of a given model.
2) Permute the values for the selected variables, then train and test the model again to calculate its new RMSE.
3) Repeat steps 1) and 2) several times to reduce bias. The average difference between the old and new RMSEs can reflect the importance of the variables. The higher the value, the more important the variable is.

5. Results

This section has been divided into two sub-sections. The first sub-section describes the calculation of coefficients in the model and the process to determine the model structure. The second sub-section demonstrates the analysis results of different machine learning methods used in the study.

5.1. Structure identification

The space–time autocorrelation coefficients and partial autocorrelation coefficients up to the 2nd-order spatial and 10th-order temporal lag are shown in Table 3 below.

Fig. 4 below provides a visual representation for the trends of the two coefficients.

As noted in Section 4.2, the specific structure of the model can be determined by the trends of space–time autocorrelation and space–time partial autocorrelation. Fig. 4 shows that, in the temporal dimension, the autocorrelation coefficient is trailing (fluctuating around 0.2), while the partial autocorrelation coefficient is subject to truncation between the 1st and 2nd order (always smaller than 0.1). However, the correlations between the spatial variables are not as clear as those between the temporal ones. One reasonable explanation of this result is that there might be correlation between the dependent variable and the 1st-order spatial autocorrelation variables. A method commonly used to incorporate more information into models of this sort involves, first, including all the variables that are possibly correlated, and then, second, examining their significance to determine the final structure. The preliminary structure of a model that includes the 1st and 2nd order of temporal lag variables and also the 1st order spatial lag variables can be defined as follows:

$$\mathbf{Y}_t = \mathrm{a}\boldsymbol{Y}_{t-1} + b\boldsymbol{Y}_{t-2} + c\boldsymbol{W}^1\boldsymbol{Y}_t + \boldsymbol{DX} + \boldsymbol{\varepsilon} \tag{7}$$

Table 3
Space-time Autocorrelation and Partial Autocorrelation Coefficients.

Temporal lag order	Spatial lag order					
	Autocorrelation			Partial Autocorrelation		
	0	1	2	0	1	2
0	1.000	0.112	0.117	–	–	–
1	0.402	0.073	0.068	0.244	0.013	0.016
2	0.290	0.042	0.032	0.068	−0.015	−0.023
3	0.254	0.039	0.033	0.047	−0.003	−0.003
4	0.242	0.038	0.032	0.040	−0.005	−0.006
5	0.251	0.042	0.030	0.045	−0.001	−0.003
6	0.249	0.042	0.037	0.040	**	**
7	0.245	0.045	0.031	0.043	**	**
8	0.248	0.041	0.038	0.005	−0.002	−0.001
9	0.232	0.042	0.034	0.003	**	**
10	0.212	0.040	0.038	0.031	−0.002	−0.006

(**: absolute value<0.001).

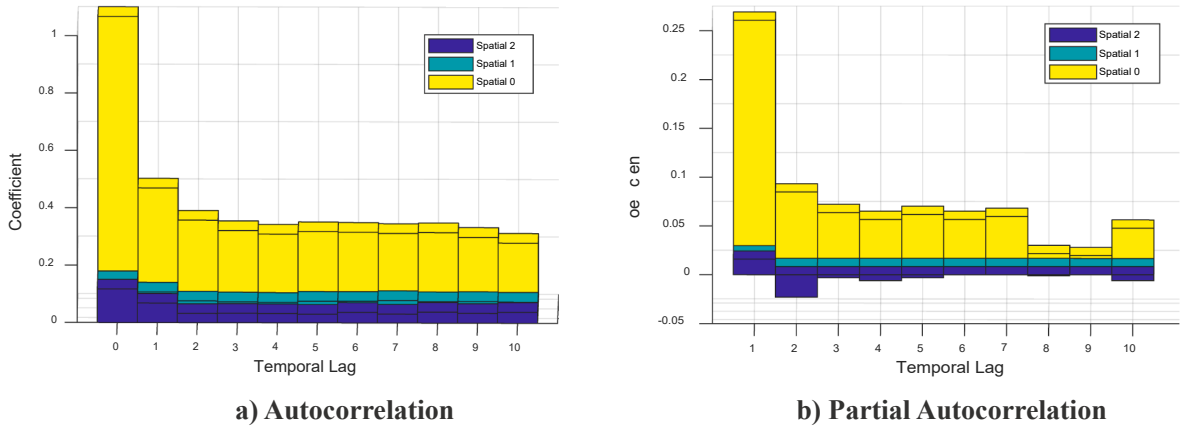


Fig. 4. Trends of Autocorrelation and Partial Autocorrelation.

Where:

$\boldsymbol{Y_t}$ A vector containing the value of the dependent variable at time t.

$\boldsymbol{Y_{t-1}/Y_{t-2}}$ A vector containing the value of the dependent variable at time t-1/t-2—this being the 1st/2nd-order temporal lag of $\boldsymbol{Y_t}$.

$\boldsymbol{W^1}$ 1st-order spatial weight matrix.

a/b/c Coefficient values of spatial and temporal variables.

D A coefficient matrix containing the coefficient values of the other independent variables discussed in section 3.3.

X A matrix containing all other independent variables.

$\boldsymbol{\varepsilon}$ A vector of random disturbance terms.

That is to say, it is believed according to the historical records that the utility condition (1 for used at least once a day while 0 for not in the model) of a target vehicle on a specific date, is linked to the utility condition of previous 1 and 2 days and the utility condition of adjacent vehicles in 0.5 km range as well as other factors. After knowing this, the approaches to predict the realistic demand of the target carsharing station has been introduced in Fig. 2 above.

5.2. Estimation results

The correlations of independent variables are initially examined by successively adding them to the model. During this process, we generated four distinct models containing different variables. These models are as follows:

Model 1: Contains information about the target vehicle and region only ($\mathbf{x^1}$-$\mathbf{x^{12}}$).

Model 2: Adds information about adjacent times and regions to Model 1 ($\mathbf{x^1}$-$\mathbf{x^{21}}$).

Model 3: Adds social media data to Model 2 ($\mathbf{x^1}$-$\mathbf{x^{23}}$).

Model 4: Adds spatial and temporal variables $\mathbf{Y_{t-1}}$, $\mathbf{Y_{t-2}}$, and $\mathbf{W^1Y_t}$ to Model 3.

The results are shown in Table 4 as follows:

Evaluation matrix of the methods when applied in ***Model 4*** are shown in Table 5 below:

In the evaluation matrix, we used five criteria to evaluate the performance of each classification model as follows:

Accuracy	Number of samples that are correctly predicted/total number.
Recall	For all the samples with the value 1, the number of samples that are correctly predicted/total number of samples.
Specificity	For all the samples with value 0, the number of samples that are correctly predicted/total number of samples.
Precision	Recall/(Recall + (1 – Specificity)).
F-Score	The harmonic mean of recall and precision: 2 * Recall * Precision/(Recall + Precision)

Compared with the parametric methods, the main advantage of the non-parametric method is they helps to explore the non-linear relationship between variables. Since most of the correlations between variables are not strictly linear, theoretically, non-parametric model such as *RF* and *XgBoost* will have a better performance in estimation accuracy, as is shown in Table 5 above.

Table 4
Accuracy of the Four Models with Different Machine Learning Method.

Model No.	*L-MLE*	*BL*	*CART*	*SVM*	*NBC*	*RF*	*XgBoost*
1	55.7%	57.2%	66.3%	71.4%	56.5%	70.8%	72.0%
2	61.1%	63.5%	69.8%	77.0%	60.3%	77.1%	79.7%
3	62.1%	65.4%	71.3%	79.7%	60.4%	79.3%	82.2%
4	63.3%	70.6%	74.9%	85.6%	62.0%	84.8%	88.1%

Table 5
The Evaluation Matrix of Model 4.

Criterion	*L-MLE*	*BL*	*CART*	*SVM*	*NBC*	*RF*	*XgBoost*
TP	25,626	28,389	29,979	34,332	24,036	33,953	35,051
FP	12,226	9463	7873	3520	13,816	3899	2801
TN	20,678	23,334	24,868	28,392	21,388	28,078	29,514
FN	14,730	12,074	10,540	7016	14,020	7330	5894
Accuracy	63.3%	70.6%	74.9%	85.6%	62.0%	84.8%	88.1%
Recall	67.7%	75.0%	79.2%	90.7%	63.5%	89.7%	92.6%
Specificity	58.4%	65.9%	70.2%	80.2%	60.4%	79.3%	83.6%
Precision	61.9%	68.7%	72.7%	82.1%	61.6%	81.2%	84.8%
F-Score	64.7%	71.7%	75.8%	86.2%	62.5%	85.3%	88.5%

However, parametric models may present whether a target variable is positive or negative correlated with the real demand. It is also the key information required for subsequent analysis.

Due to the limitation of page and table size, only the coefficients, *p-values* of the *L-MLE* and *BL* together with the OOB factor of *XgBoost* method (having highest accuracy) are reported in Table 6 as follows:

6. Analysis

This section provides the analysis according to the estimation results and the correlation of significant variables.

Table 6
Coefficients, P-values, and OOB-Factors for Independent Variables.

Variables	*MLE*		*BL*		*XgBoost*
	Coe	p-value	Coe	p-value	OOB-factor
intercept	0.18	0	−1.42	0	–
Variable Set 1					
Station capacity	−2.42	0	−13.55	0	3.32
Vehicle capacity	−23.25	0	−160.05	0	1.07
population	−0.86	0.31	−5.54	0.21	6.52
area	0.13	0.80	0.27	0.92	3.06
Commercial Area	−0.07	0.94	−1.30	0.74	5.83
Educational Area	−1.59	0	−8.97	0	3.05
Industrial Area	−0.49	0.29	−3.38	0.16	0.92
Parking Land Area	0.29	0.57	2.01	0.46	2.79
Rural Area	1.56	0	9.55	0	3.35
Transport Area	−1.14	0.03	−5.53	0.03	5.21
Audi	−0.44	0.36	−2.35	0.34	1.44
Hyundai	2.51	0.04	13.29	0.04	3.19
Kia	1.93	0.35	9.93	0.37	0.95
Toyota	−2.05	0.77	−9.11	0.82	1.17
Hatch	−1.97	0.75	−11.81	0.72	2.34
Hybrid	−0.01	0.99	0.19	0.97	0.85
MPV	−3.19	0.02	−14.31	0.05	0.93
SUV	−0.94	0.75	−5.45	0.74	1.83
Ultra	−2.1	0.09	−12.73	0.05	1.29
Van	1.21	0.68	1.74	0.91	1.45
Variable Set 2					
Nearest distance to vehicle of same type/same brand	0.93	0.59	2.05	0.77	3.33
How long the target vehicle has not been used	−0.10	0.82	−0.42	0.86	3.74
Variable Set 2(cont.)					
Usage rate for the past 30 days	94.27	0	463.55	0	9.20
Surrounding Commercial Area	2.54	0	14.69	0	1.65
Surrounding Educational Area	−0.37	0.45	−1.68	0.51	2.52
Surrounding Industrial Area	0.27	0.58	1.41	0.58	0.87
Surrounding Parking Land Area	0.58	0.21	3.23	0.18	1.77
Surrounding Rural Area	0.18	0.72	1.89	0.49	0.72
Surrounding Transport Area	−0.04	0.92	−0.49	0.83	0.88
Variable Set 3					
Number of users	4.48	0	22.49	0	5.10
Variable Set 4					
Time lag 1*Spatial Lag 0	96.20	0	429.30	0	12.35
Time lag 2*Spatial Lag 0	32.69	0	156.89	0	6.58
Time lag 0*Spatial Lag 1	21.80	0	111.30	0	8.86

6.1. Overall model analysis

As noted in section 5.2, different sets of variables were added to the model successively to form in a step-by-step manner. The aim of the process is to prove that all four sets contain relevant information regarding the car sharing operation. Table 4 reveals that the accuracy of the models improves as more variables are added. Therefore, our study suggests that following sets of variables contain relevant information about the usage of target vehicles in carsharing operations:

I. target vehicle and mesh block information
II. information about nearby vehicles and surrounding mesh blocks
III. social media data
IV. space–time correlated data

In addition, the estimation accuracy of the non-parametric methods, such as *RF* and *XgBoost* is, for each of the four models, better than parametric methods. That is because *L-MLE* and *BL* methods are essentially based on linear correlations, whereas independent variables carrying important information may not be linearly related to the dependent variables. In turn, since the variables in question provide extra information, the non-parametric methods of estimation show a better overall performance.

At the same time, however, due to the complexity of its forest structure, non-parametric methods may suffer when they come to explanatory power. As Table 5 shows, a commonly used significance indicator in the *Xgboost* method—namely, the *OOB-factor* for variables—can only point out which variables are more significant as compared to others used in the analysis. The specific relationships between dependent and independent variables prove difficult to capture, because they remain hidden in various tree structures in the forest (Prinzie and Van den Poel, 2008, Prinzie and Van den Poel, 2007). This is why a variety of different estimation methods, parametric and non-parametric, are used in the present paper. Incorporating results from the presented methods, our next section explains the correlations of each variable.

6.2. Variable analysis

Significant variables from the model have been marked with different colours in Table 5 based on their relative importance in the

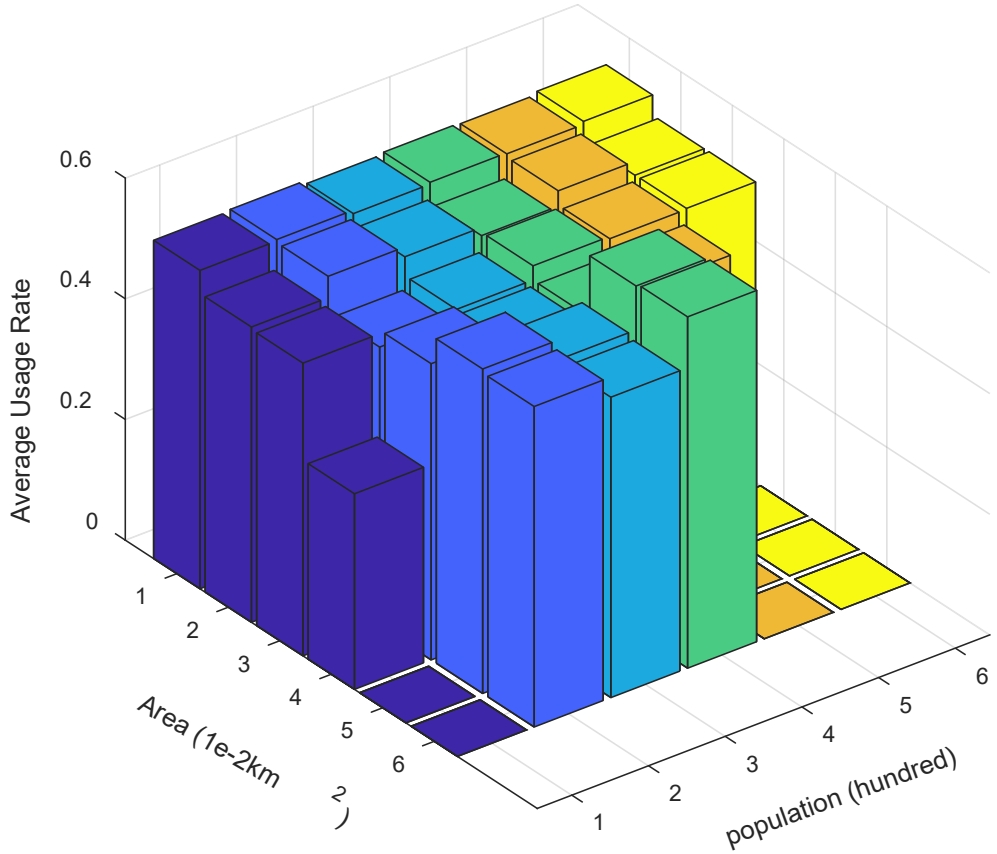


Fig. 5. Area and Population vs. Vehicle Usage Rate.

different estimation methods. Variables marked in blue show a strong linear relationship in the parametric (*L-MLE* and *BL)* methods (*p-value* < 0.05) but a relatively insignificant relationship in the *XgBoost* method (*OOB-factor* < 3). Conversely, variables marked in green are useful in classifiers (*OOB-factor* > 3) but perform poorly in the parametric estimation methods (*p-value* > 0.05). Variables marked in yellow are significant in both the parametric and non-parametric methods, while in white are significant in neither. Negative intercepts in the parametric methods demonstrate that it is rare for a station at randomly selected location to have any service demand. By analyzing the significance of the variables using all of these methods, several strategies for efficiently operating a carsharing station can be outlined.

1) **Station Capacity and Vehicle Capacity**

As mentioned in section 3, there is an outlier station which contains a large number of vehicles. However, due to the limited data provided, those vehicles cannot be excluded from the analysis. Therefore, it is inaccurate to reduce the relationship between vehicle usage and station capacity to a negative correlation. The location of stations and distribution of vehicles will be discussed further below, in connection with spatial and temporal correlation variables.

In addition, the negative correlation between usage and vehicle capacity indicates that the demand for large and medium-sized vehicles was lower than what service provider expected or could incorporate into their operation considerations. Although the overall proportion of large and medium-sized vehicles is only around 5% of all available vehicles, the lower-than-expected utilization still suggests that resources are being unutilized. Later in this section there will be a more detailed discussion of vehicle brand and type selections.

2) **Population, Area, and Land Use**

In this study, population and area are typical examples of non-linearly related independent variables with higher *OOB-factors*. Fig. 5 reveals the relationship between area and population against average usage rate.

In Fig. 5, the x-label and y-label are the mesh block area and the mesh block population where target stations located at. Mesh block system used in this figure is established by Australian statistic department. A mesh block with smaller area often appears in an urban centre or some other densely populated regions. While larger mesh blocks are used in rural region or forest land area.

Fig. 5 illustrates that the usage tends to be more frequent in the parts with higher and lower ranges, and less frequent in parts with a middle range. The regions with higher populations and larger areas suffer from lack of sufficient data. But there are three other regions (upper left corner, upper right corner and middle and lower side) that suggest different types of land use which are relevant in the context.

According to the mesh block division, high spot at the upper left corner represents the mesh block with small area and less residence where are mostly commercial or business-oriented areas in a city centre. Meanwhile, the spot at the upper right corner represents the mesh block with small areas and large amount of residence. Those mesh blocks are mostly large-sized, high-density residential areas. These two types of regions (commercial central and residential area) have always been hot spots for TNC providers. This fact has been discovered in this paper among the data.

However, the high spot at middle and lower side with larger areas is also worth discussing. In the available dataset we found that 60 vehicles have been provided by *GoGet* in these areas and their usage rate is 2%-5% higher than average. Most of these areas are suburbs situated between urban and rural areas, with combinations of residential, commercial, and rural modes of land use. Carsharing services could offer another choice for the residents in those areas, facilitating their trips between town and cities, or between different towns.

3) **Selection of Vehicle Type and Brand**

Both MPV-type and Ultra-type vehicles are strongly negative correlated with the dependent variable. A possible reason might be their uncommon capacity compared to smaller standard cars (normally with 5 seats). Their comparatively larger or smaller size may give users a sense of insecurity and unfamiliarity. Meanwhile, there is no strong correlation between usage rates and other vehicle-type variables (Hatch, SUV, Hybrid). It appears that most users prefer to get a 'normal' sized vehicle with seating for four or five passengers, but it is less sensitive about specific types of vehicles.

In the provided data, the brand variable 'Toyota' and type variable 'Hatch' (each accounted for more than 70% of its respective category) do not show a strong correlation with vehicle usage rates. The highlight of the service is 'Hyundai' brand. 39 different Hyundai vehicles are provided and the available data reveals that usage rate of Hyundai vehicles is 15% higher than average. This finding suggests that the company should increase the proportion of "Hyundai" vehicles in their service.

4) **Social Media Data**

As it has been shown 3.1–3) Social Media Data from Foursquare, the number of users and the total number of check-ins are strongly correlated with each other. To avoid multicollinearity, only one of these two variables could be included in the model. Results shows that including social media data increased the accuracy of the estimation methods by around 2% which could help providers to build more appropriate demand estimation models.

5) **Spatial and Temporal Variables**

All of the spatial and temporal variables in the model are strongly correlated with the dependent variable. That is to say, the dependent variable, whether the target vehicle has been used during the observation day, has a strong time–space autocorrelation.

In the time dimension, both the variables of time lag 1 and time lag 2 as well as the usage rate for the previous month are positively correlated with usage. This finding suggests that a target vehicle tends to be used, or else not used, continuously. Some of the available vehicles are continuously used while others were not used at all for long periods. So, according to the observation, both demand loss and resource waste obtain, simultaneously.

In the spatial dimension, the 1st-order spatial correlated variable is positively correlated with usage. This finding suggests that carsharing demand forms an aggregation state over the selected range (0.5 km). In other words, the target vehicle will tend to be used when other vehicles around are used, and vice versa. So, we recommend that the distribution of vehicles be focused more on hot spots than it is now.

6.3. Overall policy recommendations

This section provides recommendations to TNC providers (particularly *GoGet*) to design a more efficient carsharing system according to the analysis result.

According to the analysis of vehicle capacity and vehicle type in Section 6.2-1), all types of vehicles with non-standard sizes are significantly negatively related to vehicle usage, while variables associated with standard-sized vehicles do not show a significant correlation. This finding suggests that the company should reduce part of its fleet —namely, vans, buses, or ultra-type vehicles—in order to avoid under-utilized resources.

Moreover, given that users do not seem to care as much about the type of normal-sized vehicles they booked, TNC providers could consider purchasing same types of vehicles (e.g. Hatch) which allows them to reduce operating costs. Similarly, as discussed in Section 6.2-3), users show little preference for vehicle brands neither. The only exception as it was found in the available dataset is Hyundai-brand vehicles. The variable of Hyundai-brand is significantly related to usage, meaning that Hyundai is well-received by users, suggesting, in turn, the service provider could consider increasing the overall proportion of Hyundai vehicles in its fleet. In a conclusion, with respect to vehicle type and brand, the analysis shows that instead of giving more choices to users, providing a few well-received types and brands of vehicles might be a better approach to reduce maintenance costs and increase the average usage rate of vehicles.

In addition, it is worth underscoring the importance of social media for the process of policy development. By comparing the estimation results of model 2 and 3 in this paper, we find that social media data are an appropriate complementary data that can be used to improve the accuracy of the model. Thus, it is reasonable to infer that, for both existing stations and newly planned stations, a moderate level of advertising on social media is likely to improve carsharing firms' business."

Last and the most, analysis results present that there is a significant correlation between spatial and temporal information and vehicle usage. In order to generate a more efficient car sharing system and increase the usage rate of provided service to make more profits, it is recommended that TNC providers may adjust the vehicle distribution of existing stations based on spatial and temporal data. The carsharing system of *GoGet* in Sydney (the case used in the present study), for example, shows a continuous trend in time and an aggregation trend in space as discussed in Section 6.2-5). Meanwhile, there are some vehicles and stations which are far below the usage average. Both continuous trend and aggregation trend suggest that TNC providers should consider shutting down some over-serviced stations, or at least reducing the number of available vehicles at those stations. Surplus vehicles can be aggregated to the hot spots such as commercial centres or large residential areas.

For planning new stations, in addition to shedding light on users' selection of vehicle types and brands, the model also illuminates the strategy of location selection. Based on the analysis of land-use and socio-demographic variables, suburbs in semi-urban areas represent an under-exploited market. Vehicles at existing stations in those areas have higher usage rate, and companies could thus consider increasing service in such areas, particularly if the areas lack public transport or similar carsharing options. Meanwhile, hot-spot areas such as commercial centres and large-scale residential areas should be focused as well, since, in the model, variables pertaining to land-use characteristics are significantly correlated with vehicle usage. After the location is selected, the real demand for the station can be estimated by collecting land-use and socio-demographic data and the usage condition of vehicles at nearby stations. Then *STARMA* model similar to the one outlined in this paper could be applied.

7. Conclusion

This paper explores how to estimate the actual demand for a carsharing station based on vehicle-renting records, among other types of data. The critical contributions of the study can be concluded as:

- The study is the first attempt at carsharing realistic demand estimation by exploring latent information from the historical record with missing parts and a large set of other seemingly unrelated data sources.
- The study expands the scope of spatio-temporal autocorrelation analysis on the new, first-discussed problem to solve the problem of demand under-estimating caused by the missing data in historical records.
- The study attempts to improve the estimation accuracy of the model developed for this purpose by adding emerging data sources, including social media data and spatio-temporal correlated data.
- The study develops a correlation analysis of significant variables and provides suggestions accordingly to the TNC's operation policy.

In light of our findings, the study provides suggestions to the service providers concerning their operation policies. Generally speaking, to improve efficiency, the distribution of available vehicles can be more aggregated in 'hot spots'. Also, regarding vehicle types, more mid-sized vehicles with seats for four or five passengers should be provided, in favour of small or large-sized vehicles. In addition, new stations can be optimally located at emerging central areas within cities, newly built large-scale residential areas, or towns situated between urban and rural areas.

In future research, other spatio-temporal related variables can be introduced for the purpose of further developing an estimation model for real carsharing demand. A variable worth considering is carsharing vehicle trajectory. Most carsharing companies allow users to rent and return vehicles at different stations. We hypothesize that the origin and destination stations of the shared vehicles—this no doubt being a spatio-temporal correlation variable as well—also carries a large amount of latent information. Factoring this additional information into the analysis should help further improve the accuracy of the model. Meanwhile, applying novel machine learning and deep learning approaches on coefficient calculation of spatio-temporal autocorrelation model and select the most suitable one for the problem is also worth to be discussed.

CRediT authorship contribution statement

Zesheng Cheng: Conceptualization, Data curation, Formal analysis, Methodology. **Taha Hossein Rashidi:** Conceptualization, Methodology, Investigation, Project administration, Resources, Supervision, Writing – review & editing. **Sisi Jian:** Conceptualization, Methodology, Formal analysis, Supervision, Writing – review & editing. **Mojtaba Maghrebi:** Supervision, Writing – review & editing. **Steven Travis Waller:** Project administration, Resources, Supervision, Writing – review & editing. **Vinayak Dixit:** Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

We have tried to apply functionality similarity matrix (F) and mobility pattern similarity matrix (M) in the model. Below is comparison of the regression accuracy with and without the two matrices.

	***L-MLE* method**	***BL***	**Xgboost**
Without F & M	63.3%	70.6%	88.1%
With F & M	65.3%	71.2%	86.6%

It can be noticed that there is a small improvement in the accuracy of the *L-MLE* and the *Binary Logistic* models, while as for the *XgBoost* methods, the accuracy of the model roughly remain unchanged. Accordingly, for the models with similarity matrices, the information carried by the two similarity matrices (M & F) might be covered by other variables in the model already. For example, the information provided by the functionality similarity matrix might be covered by the land use variables (variable x3-x8 in the paper). So we do not report the accuracy table and the two similarity matrices in the main sections of the paper.

In other cases, similarity matrices might be better options for problems which have a part of data missing. When calculating similarity, only variables which have data existing for both terms will be considered while other variables will be ignored or specially treated. In addition, especially for functionality similarity matrix, it contains a large amount of term 1 (which means the two vehicles have same characteristics) when taking single vehicle as research objective. That is because some of the provided vehicles has same stations and then same land use characteristics. Therefore, similarity matrices might be more suitable for those problems which research object is carsharing station instead of single vehicles.

References

Boesch, P.M., Ciari, F., Axhausen, K.W., 2016. Autonomous vehicle fleet sizes required to serve different levels of demand. Transp. Res. Rec. 2542, 111–119.
Breiman, L, Ihaka, R, 1984. Nonlinear discriminant analysis via scaling and ACE[M]. Davis One Shields Avenue, Davis, CA, USA.
Celsor, C., Millard-Ball, A., 2007. Where does carsharing work? Using geographic information systems to assess market potential. Transp. Res. Rec. 1992, 61–69.
Cepolina, E.M., Farina, A., 2012. A new shared vehicle system for urban areas. Transp. Res. Part C: Emerging Technol. 21, 230–243.
Cervero, R., Golub, A., Nee, B., 2007. City CarShare: longer-term travel demand and car ownership impacts. Transp. Res. Rec. 1992, 70–80.
Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., 2015. Xgboost: extreme gradient boosting. R package version 0.4-2, 1, 1–4.
Ciari, F., Schuessler, N., Axhausen, K.W., 2013. Estimation of carsharing demand using an activity-based microsimulation approach: model discussion and some results. Int. J. Sustainable Transp. 7, 70–84.
Cliff, A., Ord, J.K., 1975. Space-time modelling with an application to regional forecasting. Trans. Institute Brit. Geogr. 119–128.
Cliff, A D, Ord, J K, 1981. Spatial processes: models & applications[M]. Taylor & Francis.

Codecá, L., Frank, R., Faye, S., Engel, T., 2017. Luxembourg sumo traffic (lust) scenario: Traffic demand evaluation. IEEE Intell. Transp. Syst. Mag. 9, 52–63.
Costain, C., Ardron, C., Habib, K.N., 2012. Synopsis of users' behaviour of a carsharing program: a case study in Toronto. Transp. Res. Part A: Policy Practice 46, 421–434.
Cressie, N., Majure, J.J., 1997. Spatio-temporal statistical modeling of livestock waste in streams. J. Agric. Biol. Environ. Statistics 24–47.
Curry, I., 1970. Univariate spatial forecasting. Econ. Geogr. 46, 241–258.
de Lorimier, A., El-Geneidy, A.M., 2013. Understanding the factors affecting vehicle usage and availability in carsharing networks: a case study of Communauto carsharing system from Montréal, Canada. Int. J. Sustainable Transp. 7, 35–51.
Deutsch, S.J., Ramos, J.A., 1986. Space-time modeling of vector hydrologic sequenceS 1. JAWRA J. Am. Water Resources Assoc. 22, 967–981.
Dong, C., Lian, C., Hu, S., Deng, Z., Gong, J., Li, M., Liu, H., Xing, M., Zhang, J., 2018. Size-dependent activity and selectivity of carbon dioxide photocatalytic reduction over platinum nanoparticles. Nat. Commun. 9, 1–11.
El-Assi, W., Salah Mahmoud, M., Nurul Habib, K., 2017. Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. Transportation 44 (3), 589–613.
Engel, J., 1988. Polytomous logistic regression[J]. Statistica Neerlandica 42 (4), 233–252.
Fagnant, D.J., Kockelman, K.M., 2014. The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. Transp. Res. Part C: Emerging Technol. 40, 1–13.
Firnkorn, J., Müller, M., 2011. What will be the environmental effects of new free-floating car-sharing systems? The case of car2go in Ulm. Ecol. Econ. 70, 1519–1528.
Habib, K.M.N., Morency, C., Islam, M.T., Grasset, V., 2012. Modelling users' behaviour of a carsharing program: Application of a joint hazard and zero inflated dynamic ordered probability model. Transp. Res. Part A: Policy Practice 46, 241–254.
Zhang, X, Liu, W, Waller, S T, et al., 2019. Modelling and managing the integrated morning-evening commuting and parking patterns under the fully autonomous vehicle environment[J]. Transportation Research Part B: Methodological 128, 380–407.
Hastie, T., Tibshirani, R., Friedman, J. 2009. Overview of supervised learning. The elements of statistical learning. Springer.
Jian, S., Rashidi, T.H., Dixit, V., 2017. An analysis of carsharing vehicle choice and utilization patterns using multiple discrete-continuous extreme value (MDCEV) models. Transp. Res. Part A: Policy Practice 103, 362–376.
Jian, S., Rashidi, T.H., Wijayaratna, K.P., Dixit, V.V., 2016. A Spatial Hazard-Based analysis for modelling vehicle selection in station-based carsharing systems. Transp. Res. Part C: Emerging Technol. 72, 130–142.
Jorge, D., Correia, G., 2013. Carsharing systems demand estimation and defined operations: a literature review. Eur. J. Transport Infrastruct. Res. 13.
Kamarianakis, Y., Prastacos, P., 2005. Space–time modeling of traffic flow. Comput. Geosci. 31, 119–133.
Leung, K.M., 2007. Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007, 123-156.
Lin, J., 2017. *Realistic* demand and route choice of archives management in Colleges and universities in the information age. In: 2017 4th International Conference on Education, Management and Computing Technology (ICEMCT 2017), 2017. Atlantis Press, 85-89.
Lv, Z, Li, J, Dong, C, et al., 2021. Deep learning in the COVID-19 epidemic: A deep model for urban traffic revitalization index[J]. Data & Knowledge Engineering 135, 101912.
Lv, Z, Li, J, Dong, C, DeepSTF, et al., 2021. A deep spatial–temporal forecast model of taxi flow[J]. The Computer Journal.
Martin, R.L., Oeppen, J., 1975. The identification of regional forecasting models using space: time correlation functions. Trans. Institute Brit. Geogr. 95–118.
Morency, C., Habib, K.M.N., Grasset, V., Islam, M.T., 2012. Understanding members' carsharing (activity) persistency by using econometric model. J. Adv. Transp. 46, 26–38.
Morency, C., Trepanier, M., Agard, B., 2011. Typology of carsharing members. Transp. Res. Board 90th Annual Meeting 1236.
Noble, W.S., 2006. What is a support vector machine? Nat. Biotechnol. 24, 1565–1567.
Pfeifer, P.E., Deutrch, S.J., 1980. A three-stage iterative procedure for space-time modeling phillip. Technometrics 22, 35–47.
Pfeifer, P.E., Deutsch, S.J., 1980. A STARIMA model-building procedure with application to description and regional forecasting. Trans. Institute Brit. Geogr. 330–349.
Prinzie, A., van den Poel, D., 2007. Random multiclass *classification*: Generalizing random forests to random mnl and random nb. In: International Conference on Database and Expert Systems Applications, Springer, 349-358.
Prinzie, A., van den Poel, D., 2008. Random forests for multiclass classification: Random multinomial logit. Expert Syst. Appl. 34, 1721–1732.
Schaefers, T., 2013. Exploring carsharing usage motives: a hierarchical means-end chain analysis. Transp. Res. Part A: Policy Practice 47, 69–77.
Schmöller, S., Weikl, S., Müller, J., Bogenberger, K., 2015. Empirical analysis of free-floating carsharing usage: the Munich and Berlin case. Transp. Res. Part C: Emerging Technol. 56, 34–51.
Sedgwick, P., 2014. Spearman's rank correlation coefficient. BMJ 349, g7327.
Statistics, A. B. O. 2016. *Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas* [Online]. Available: https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1270.0.55.001~July%202016~Main%20Features~Mesh%20Blocks%20(MB)~10012 [Accessed 30 April 2020].
STATISTICS, A. B. O. 2016. *Census of Population and Housing: Mesh Block Counts, Australia, 2016* [Online]. Available: https://www.abs.gov.au/ausstats/abs@.nsf/mf/2074.0 [Accessed 30 April 2020].
Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46, 234–240.
Yao, X., 2003. Research issues in spatio-temporal data mining. Workshop on geospatial visualization and knowledge discovery, University Consortium for Geographic Information Science, Virginia, 1-6.